Refining the impact of genetic evidence on clinical success

Eric Vallabh Minikel[1], Jeffery L Painter[2,*], Coco Chengliang Dong[3], Matthew R. Nelson[3,4†]

1. Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA 02142
2. JiveCast, Raleigh, NC, 27601
3. Deerfield Management Company, L.P., New York, NY 10010
4. Genscience LLC, New York, NY 10010

*Present address: GlaxoSmithKline, Research Triangle Park, NC 27709
†Corresponding author: mnelson@genscience.com

Abstract

The cost of drug discovery and development is driven primarily by failure, with just ~10% of clinical programs eventually receiving approval. We previously estimated that human genetic evidence doubles the success rate from clinical development to approval. In this study we leverage the growth in genetic evidence over the past decade to better understand the characteristics that distinguish clinical success and failure. Relative success (RS) between drug mechanisms with genetic support and those without varies significantly among therapy areas and development phases and improves with increasing confidence in the causal gene. RS is largely unaffected by genetic effect size, minor allele frequency, or year of discovery. Our findings suggest that we are far from reaching peak genetic insights to aid the discovery of targets for more effective drug therapies.

The cost of drug discovery and development is driven primarily by failure[1], with just ~10% of clinical programs eventually receiving approval[2–4]. We previously estimated that human genetic evidence doubles the success rate from clinical development to approval[5]. In this study we leverage the growth in genetic evidence over the past decade to better understand the characteristics that distinguish clinical success and failure. Relative success (RS) between drug mechanisms with genetic support and those without varies significantly among therapy areas and development phases and improves with increasing confidence in the causal gene. RS is largely unaffected by genetic effect size, minor allele frequency, or year of discovery. Our findings suggest that we are far from reaching peak genetic insights to aid the discovery of targets for more effective drug therapies.

Human genetics is one of the only forms of scientific evidence that can demonstrate the causal role of genes in human disease, providing a crucial tool for identifying and prioritizing potential drug targets and providing insights into the effect of pharmacological engagement, dose-response relationships[6,7] and safety risks[8]. Nonetheless, many questions remain about the application of human genetics in drug discovery. Genome-wide association studies (GWAS) of common, complex traits, including diseases, generally identify variants of small effect. This contributed to early skepticism of the value of GWAS[9]. Anecdotally, such variants can point to highly successful drug targets[6,7], and yet, genetic support from GWAS is less predictive of drug target advancement than support from Mendelian disease[5,10].

In this paper we investigate several open questions regarding the use of genetic evidence for prioritizing drug discovery. We explore the characteristics of genetic associations that are more likely to differentiate successful from unsuccessful drug mechanisms, exploring how they differ

across therapy areas and among discovery and development phases. We also investigate how close we are to saturating the insights we can get from genetic studies for drug discovery and how much of the genetically supported drug discovery space remains unexplored.

After filtering Citeline Pharmaprojects for monotherapy programs added since 2000 annotated with a highest phase reached and assigned a human gene target, we obtained 29,476 target-indication (T-I) pairs for analysis (Fig. S1). Intersecting with 79,095 unique human gene-trait (G-T) pairs (see Methods) yielded an overlap of 2,153 T-I and G-T pairs (7.3%) with an indication-trait similarity ≥0.8 (Fig. S1, S2A). The probability of having genetic support was higher for launched T-I pairs than those in historical or active clinical development (Figure 1A), consistent with our previous findings[5,10]. We defined relative success (RS) as the ratio of the probability of success with genetic support to the probability of success without genetic support. Focusing on historical programs — those with known outcomes of either discontinuation or launch — we tested the sensitivity of RS to various characteristics of genetic evidence. RS was sensitive to the indication-trait similarity threshold (Figure S2A), which we set to 0.8 for all analyses herein. RS was >2 for all sources of human genetic evidence examined (Figure 1B). The highest value for OMIM (RS = 3.4) was not the result of a higher success rate for orphan drug programs (Figure S2B), a designation commonly acquired for rare diseases. Rather it may be due to the difference in confidence in causal gene assignment between Mendelian conditions and GWAS. The RS for Open Targets GWAS (OTG) associations was sensitive to the confidence in variant-to-gene mapping as reflected in the minimum share of locus-to-gene score (Figure 1C). Somatic evidence from IntOGen had an RS of 2.4 in oncology (Figure S2C), similar to GWAS overall. Analyses below are limited to germline genetic evidence unless otherwise noted.

As sample sizes grow ever larger with a corresponding increase in the number of unique G-T associations, some expect[11] the value of GWAS genetic findings to become less useful for the purpose of drug target selection. We explored this in several ways. We investigated the year that a genetic support for a T-I pair was first discovered, under the expectation that more common and larger effects are discovered earlier. Although there was a slightly higher RS for discoveries from 2007-2010 that was largely driven by early lipid and cardiovascular-related associations, the effect of year was not significant (P = 0.37, Figure 1D). Results were similar when replications or OMIM discoveries were included (Figure S2D-E). We next divided up GWAS associations by the number of unique G-T associations. Counter to our expectations, probability of launch increased with the number of associated genes by 0.045% per gene (P = 0.031, Figure 1D). There were no statistically significant associations with estimated effect sizes (P = 0.93 and 0.59, for quantitative and binary traits, respectively) nor minor allele frequency (P = 0.43, Figure 1D). We illustrate that ever larger GWAS can continue to uncover support for successful targets with a recent large GWAS that more than doubled the number of significantly associated loci with type 2 diabetes[12]: of 12 launched drug mechanisms, the number with GWAS support increased from 5 to 7 (Figure S3).

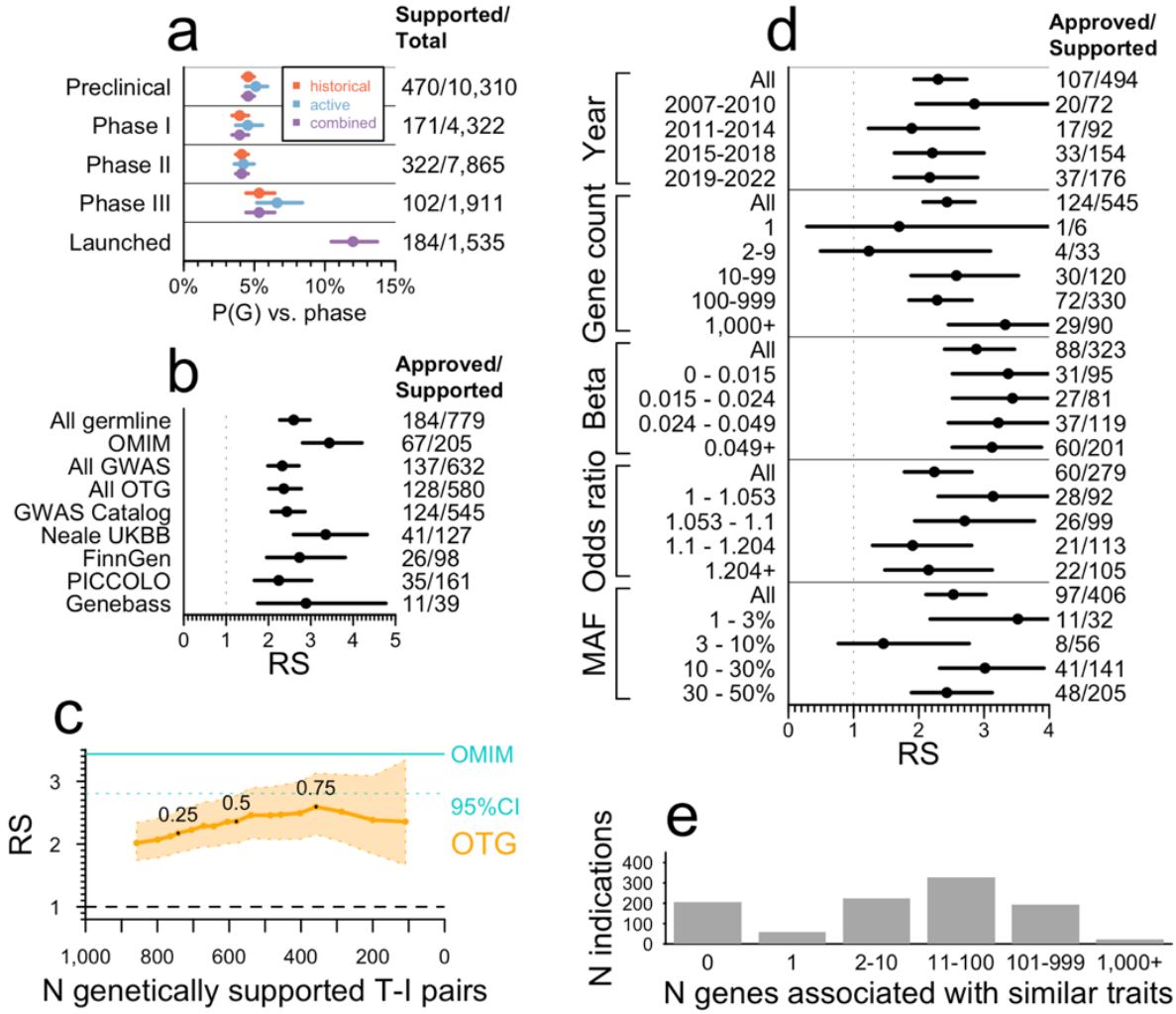


Figure 1. Impact of genetic evidence characteristics on relative success. A) *Proportion of target-indication (T-I) pairs with genetic support, P(G), as a function of highest phase reached. Bars are Wilson 95% confidence intervals.* ***B)*** *Sensitivity of relative success (RS) from phase I–launch of T-I pairs with genetic evidence to source of human genetic association. Bars are Katz 95% confidence intervals.* ***C)*** *Sensitivity of RS to locus-to-gene (L2G) share threshold among Open Targets Genetics (OTG) genome-wide association study (GWAS) significant associations. The minimum L2G share required for inclusion in the dataset is varied from 0.1 to 1.0 in increments of 0.05 (labels) and RS (y axis) is plotted against the number of clinical (phase I+) programs considered to have genetic support from OTG (x axis). Shaded areas are Katz 95% confidence intervals.* ***D)*** *Sensitivity of RS for OTG GWAS-supported T-I pairs to binned variables: i) year in which a T-I pair first acquired human genetic support from GWAS, excluding replications and excluding T-I pairs otherwise supported by OMIM, ii) number of genes exhibiting genetic association to the same trait, iii) quartile of effect size (beta) for quantitative traits, iv) quartile of effect size (odds ratio, OR) for case/control traits standardized to be >1 (i.e., 1/OR if <1), and v) order of magnitude of minor allele frequency bins. Bars are Katz 95% confidence intervals.* ***E)*** *Count of indications ever in development in Pharmaprojects (y axis) by the number of genes associated with traits similar to those indications (x axis). See Figure S6 for the same analyses restricted to drugs with a single known target.*

Previously, we observed significant heterogeneity amongst therapy areas in the fraction of approved drug mechanisms with genetic support, but did not investigate the impact on probability of success[5]. Here, our estimates of RS from phase I to launch show significant heterogeneity (P < 1.0e-15), with nearly all therapy areas having estimates greater than one, 11 of 17 greater than two, and hematology, metabolic, and respiratory greater than three (Fig. 2A-E). In most therapy areas, the impact of genetic evidence was most pronounced in phases II and III and least impactful in phase I, corresponding to their capacity to demonstrate clinical efficacy. We also found evidence that genetic evidence differentiated likelihood to progress from preclinical to clinical development. This was most notable for metabolic diseases, that may reflect preclinical models that are more predictive of clinical outcomes. Probability of genetic support by therapy area was correlated with probability of success (r = 0.52, P = 0.032) and with RS (r = 0.66, n = 68, P = 0.0026; Fig. S4), which led us to explore how the sheer quantity of genetic evidence available within therapy areas (Fig. 2F, Fig. S5A) may influence this. We found that therapy areas with more possible gene-indication (G-I) pairs supported by genetic evidence had significantly higher RS (r = 0.73, P = 0.00089, Fig. 2G), although respiratory and endocrine were notable outliers with high RS despite fewer associations.

We hypothesized that genetic support might be most pronounced for drug mechanisms with disease-modifying effects, as opposed to those that manage symptoms, and that the proportion of such drugs differ by therapy area[13,14]. We were unable to find data with these descriptions available for a large number of drug mechanisms, but we reasoned that targets of disease-modifying drugs are more likely to be specific to a disease, whereas targets of symptom-managing drugs are more likely to be applied across many indications. We therefore examined the number and diversity of all-time launched indications per target. Launched T-I pairs are heavily skewed towards a few targets. Of 450 launched targets, the 42 with ≥10 launched indications comprise 713 (39%) of 1,806 launched T-I pairs (Fig. 2H). Many of these are used across diverse indications for management of symptoms such as inflammatory and immune responses (*NR3C1*, *IFNAR2*), pain (*PTGS2*, *OPRM1*), mood (*SLC6A4*), or parasympathetic response (*CHRM3*). The count of launched indications was inversely correlated with the mean similarity of those indications (ρ = -0.72, P = 1.7e-85; Fig. 2H). Among T-I pairs, the probability of having genetic support increased as the number of approved indications decreased (P = 1.8e-6) and as the similarity of a target's approved indications increased (P = 3.0e-6, Figure 2I). We observed a corresponding impact on RS, increasing in therapy areas where the similarity among approved indications increased, and decreasing with increasing indications per target (r = 0.68, P = 0.002, and r = -0.56, P = 0.023, respectively, Fig. 2J-K).

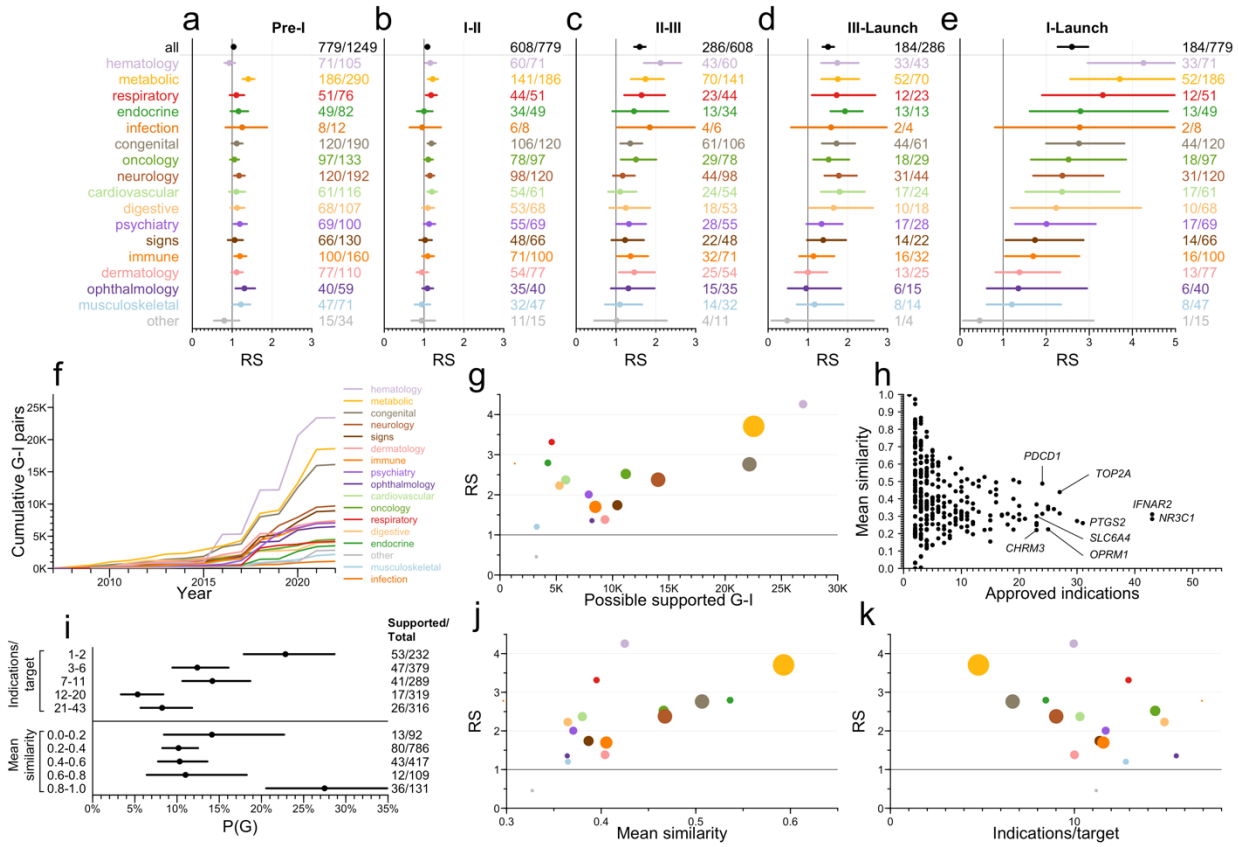


Figure 2. Differences in relative success between therapy areas and the number and diversity of indications per target. A-E) *RS by therapy area and phase transition. Bars are Katz 95% confidence intervals.* ***F)*** *Number of approved indications vs. similarity of those indications, by approved drug target.* ***I)*** *Proportion of approved target-indication pairs with genetic support, P(G), binned by quintile of the number of approved indications per target (top panel) or by mean similarity among approved indications (bottom panel). Targets with exactly 1 approved indication (6.2% of launched T-I pairs) are considered to have mean similarity of 1.0. Bars are Wilson 95% confidence intervals. G) RS (y axis) by number of supported G-I pairs (x axis) across therapy areas, dots colored as in panels A-E and sized according to number of genetically supported T-I pairs in at least phase I.* ***H)*** *Mean similarity (y axis) among approved indications per target (x axis). J) As G, with mean similarity (x axis).* ***K)*** *As G, with mean count of approved indications per target (x axis). See Figure 7 for the same analyses restricted to drugs with a single known target.*

Only 4.7% (284/6,067) of T-I pairs active in phase I-III possess human germline genetic support (Figure 1A), similar to T-I pairs no longer in development (4.2%, 595/14,098), a difference that was not statistically significant (P = 0.14). We estimated (see Methods) that only 1.1% of all genetically supported G-I relationships have been explored clinically (Figure 3A), or 2.2% when restricting to the most similar indication. Given that the vast majority of proteins are classically "undruggable", we explored the proportion of genetically supported G-I pairs that had been developed to at least phase I, as a function of therapy area across several classes of tractability[15] (Fig. 3A). Within therapy areas, oncology kinases with germline evidence were the most saturated at 100 of 230 (43%) of all genetically supported G-I pairs had reached at least

phase I, followed by nuclear receptors for musculoskeletal disorders (2/7, 29%) and GPCRs for psychiatric indications (15/53, 28%). Grouping by target rather than G-I pair, 3.7% of genetically supported targets have been pursued for any genetically supported indication (Figure S5B). Of possible genetically supported G-I pairs, most (74%) arose from OTG associations, mostly within the past 5 years (Figure 2F). Such low utilization is partly due to recent emergence of most genetic evidence (Fig. S2F-G, Fig. S5A). Because some types of targets may be more readily tractable by antagonists than agonists, we also grouped by target and examined human genetic evidence by direction of effect, for IntOGen tumor suppressors versus oncogenes (Figure 3B), identifying a few substrata for which a majority of genetically supported targets had been pursued to at least phase I for at least one genetically supported indication. Oncogene kinases received the most attention, with 19/26 (73%) reaching phase I.

To focus on demonstrably druggable proteins, we further restricted the analysis to targets with both i) any program reaching phase I, and ii) ≥1 genetically supported indication. Out of 1,130 qualifying targets, only 371 (33%) targets had been pursued for one or more supported indications (Figure 3C). Most of these targets were pursued for indications with and without genetic support (303, 27%), though most development effort has been for unsupported indications at a 17:1 ratio. Within this subset of targets, we asked whether genetic support was predictive of which indications would advance the furthest. Grouping active and historical programs by D-I pair, we found that the odds of advancing to a later stage in the pipeline is 80% higher for indications with genetic support (P = 4.8e-72, Figure 3D).

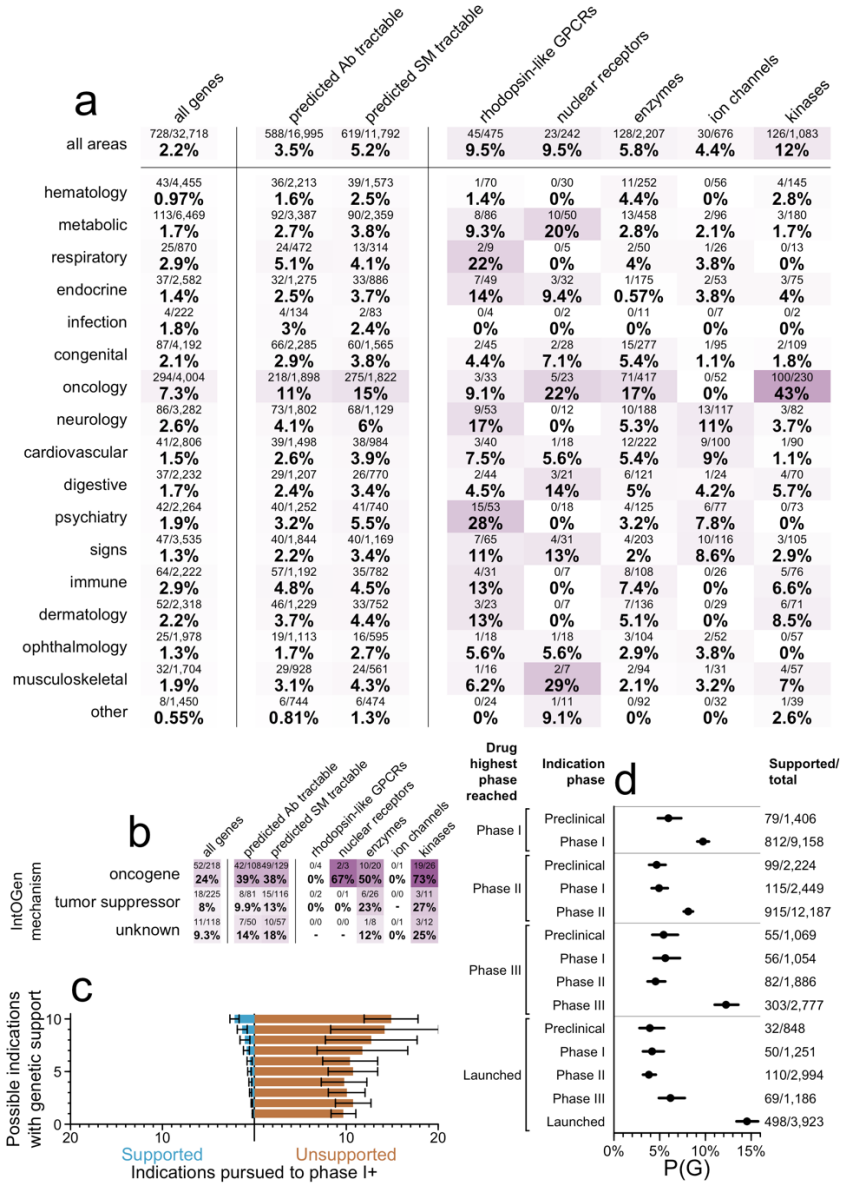


Figure 3. Clinical investigation of drug mechanisms with genetic evidence. A) *Heatmap of proportion of genetically supported T-I pairs that have been developed to at least phase I, by therapy area (y axis) and gene list (x axis).* ***B)*** *As panel A, but for genetic support from IntOGen rather than germline sources and grouped by the direction of effect of the gene according to IntOGen (y axis), and also grouped by target rather than T-I pair. Thus, the denominator for each cell is the number of targets with at least one genetically supported indication, and each target counts in the numerator if at least one genetically supported indication has reached phase I.* ***C)*** *Of targets that have both reached phase I for any indication, and have at least one genetically supported indication, the mean count (x axis) of genetically supported (left) and unsupported (right) indications pursued, binned by the number of possible genetically supported indications (y axis). Bars are Wilson 95% confidence intervals.* ***D)*** *Proportion of D-I pairs with genetic support, P(G) (x axis), as a function of the drug-indication pair's phase reached (inner y axis grouping) and the drug's highest phase reached for any indication (outer y axis grouping). Bars are Wilson 95% confidence intervals. See Figure S8 the same analyses restricted to drugs with a single known target.*

While there have been anecdotes such as *HMGCR* to argue that genetic effect size may not matter in prioritizing drug targets, here we provide systematic evidence that small effect size, recent year of discovery, increasing number of genes identified, or higher associated allele frequency do not diminish the value of GWAS evidence. Without refuting the omnigenic model, this suggests that at current sample sizes, GWAS results can help identify drug targets most likely to succeed. These results argue for continuing investment to expand GWAS-like evidence, particularly for many complex diseases with treatment options that fail to modify disease. Although genetic evidence has value across most therapy areas, its benefit is more pronounced in some areas than others. Furthermore, it is possible that the therapy areas where genetic evidence had a lower impact have seen more focus on symptom management. If so, we would predict that for drugs aimed at disease modification, human genetics should ultimately prove highly valuable across therapy areas.

The focus of this work has been on the relative success of drug programs with and without genetic evidence. We have mostly demonstrated several factors that do not impact this value. Given drug development interest within a therapy area, we have not demonstrated characteristics of the genetic evidence, aside from confidence in the causal gene, that might differentiate which genes to focus on as potential drug targets. As GWAS sample size and diversity continue to grow, the omnigenic model predicts that for complex traits most genes will have a measurable effect, even if relatively few are core to the etiology. Under this model, the vast majority of associated genes are peripheral, having effects mediated through their influence on core genes[16]. If core genes are most relevant as therapeutic drug targets, then a decreasing fraction of genes identified through GWAS over time should be considered potential drug targets. More work is required to better understand the challenges of target identification and prioritization given the genetic evidence precondition.

The utility of human genetic evidence has had firm theoretical and empirical footing for several years[5,6,10]. If the benefit of this evidence were canceled out by competitive crowding[17], then currently active clinical phases should have higher rates of genetic support than their corresponding historical phases, and might look similar to, or even higher than, approved pairs. Instead, we find that active programs continue to have low rates of genetic support, similar to historical pipelines, suggesting that human genetic data have not yet begun to appreciably influence pipeline composition across the industry. Meanwhile, only a tiny fraction of classically druggable genetically supported G-I pairs have been pursued even among targets with clinical development reported. Human genetics thus represents a growing opportunity for novel target selection and improving indication selection for existing drugs and drug candidates. Increasing emphasis on drug mechanisms with supporting genetic evidence is expected to increase success rates and lower the cost of drug discovery and development.

Methods

Drug development pipeline. Citeline Pharmaprojects[18] was queried via API (Dec 22, 2022) to obtain information on drugs, targets, indications, phases reached, and current development status (active versus historical). We removed combination therapies, diagnostic indication, and programs with no human target or no indication assigned. For most analyses, only programs added to the database since 2000 were included, while for the count and similarity of launched indications per target, we used all launches for all time. Indications were considered to have genetic insight if they had ≥0.8 similarity to i) an OMIM or IntOGen disease, or ii) a GWAS trait with at least 3 independently associated loci, based on lead SNP positions rounded to the nearest 1 Mb. For calculating relative success, we used the number of T-I pairs with genetic insight as the denominator. Many drugs had more than one target assigned, in which case all targets were retained for target-indication pair analyses. Re-running all analyses restricted to only drugs with exactly one target assigned yielded substantially similar results (Figures S5-7).

OMIM. The OMIM Gene Map (downloaded Sep 29, 2021) contained 8,246 unique gene-phenotype links. We restricted to entries with phenotype mapping code 3 ("the molecular basis for the disorder is known; a mutation has been found in the gene") and removed phenotypes with no MIM number assigned. We used regular expression matching to further filter out phenotypes containing the terms "somatic", "susceptibility", or "response" (drug response associations) and those flagged as questionable ("?"), or representing non-disease phenotypes ("["). A set of OMIM phenotypes are flagged as denoting susceptibility rather than causation ("{"); this category includes low-penetrance or high allele frequency association assertions that we wished to exclude, but also germline heterozygous loss-of-function mutations in tumor suppressor genes, where the underlying mechanism of disease initiation is loss of heterozygosity, which we wished to include. We therefore also filtered out phenotypes containing "{" except for those that did contain the terms "cancer", "neoplasm", "tumor", or "malignant" and did not contain the term "somatic". Remaining entries were evaluated for validity and disease mechanism by two curators, and gene-disease combinations for which a disease association was deemed not to have been established were excluded from all analyses. All of the above filters left 5,450 gene-trait links. MeSH terms for OMIM phenotypes were then mapped using the EFO OWL database using an approach previously described[19], with additional mappings from Orphanet, full text matches to the full MeSH vocabulary, or finally, manual curation, for a cumulative 100% mapping rate.

Open Targets Genetics. Open Targets Genetics (OTG, version 8) variant-to-disease (V2D), locus-to-gene (L2G), variant index, and study index data were downloaded from EBI. Traits with multiple EFO IDs were excluded as these generally represent conditional, epistasis, or other complex phenotypes that would lack mappings in the MeSH vocabulary. Of the top 100 traits with the greatest number of genes mapped, we excluded 76 as having no clear disease relevance (example: "red cell distribution width") or no obvious marginal value (example: excluded "trunk predicted mass" because "body mass index" was already included). Remaining traits were mapped to MeSH using the EFO OWL database, full text queries to the MeSH API, mappings already manually curated in PICCOLO (see below) or new manual curation. In total, 25,124/49,599 unique traits (51%) were successfully mapped to a MeSH ID. We included associations with $P < 5e\text{-}8$. For gene mapping we defined L2G share as the proportion of the total L2G score assigned each gene among all potentially causal genes with any L2G score. In sensitivity analyses we considered L2G and V2G share thresholds from 10% to 100% (Figure 1B and S3A), but main analyses used only genes with ≥50% L2G share (which are also the top-ranked genes for their respective associations).

PICCOLO. PICCOLO[20] tests for colocalization without full summary statistics by using Probabilistic Identification of Causal SNPs (PICS) and a reference dataset of SNP linkage disequilibrium values. We included hits with GWAS *P* < 5e-8, eQTL *P* < 1e-5, and H4 ≥ 0.9.

Genebass. Genebass[21] data from 394,841 UK Biobank participants (the "500K" release) were queried using Hail (July 22, 2022) We used SKAT pLoF gene burden tests with P < 1e-5. Because the traits in Genebass are from UK Biobank, which is included in OTG, we used the OTG MeSH mappings established above.

IntOGen. IntOGen (Feb 1, 2020) assigns each gene a mechanism in each tumor type; occasionally a gene will be classified as a tumor suppressor in one type and an oncogene in another. We grouped by gene and assigned each gene its modal classification across cancers. MeSH mappings were curated manually.

MeSH term similarity. MeSH term Lin and Resnik similarities were computed as described[22,23]. Similarities of -1, indicating infinite distance between two concepts, were assigned as 0. The two scores were regressed against each other across all term pairs, and the Resnik scores were adjusted by a multiplier such that both scores had a range from 0 to 1 and their regression had a slope of 1. The two scores were then averaged to obtain a combined similarity score. Similarity scores were successfully calculated for 1,022/1,033 (98.9%) unique MeSH terms for Pharmaprojects indications, corresponding to 99.6% of Pharmaprojects T-I pairs, and for 2,299/2,638 (87.1%) unique MeSH terms for genetic associations, corresponding to 99.8% of associations.

Therapeutic areas. MeSH terms for Pharmaprojects indications were mapped onto 16 top-level headings plus an "other". Many MeSH terms map to >1 tree position; these multiples were retained and counted towards each therapy area, except for the following conditions: for terms mapped to oncology, we deleted their mappings to all other areas; and "other" was used only for terms that mapped to no other areas.

Analysis of Vujkovic et al. 2020 type 2 diabetes GWAS. We considered as novel any genes with a novel nearest gene, novel coding variant, or a novel lead SNP colocalized with an eQTL with H4 ≥ 0.9. Non-novel nearest genes, coding variants, and colocalized lead SNPs were considered established variants. Together, these approaches identified 217 established GWAS genes and 469 novel ones. We included 19 genes from OMIM linked to Mendelian forms of diabetes or syndromes with diabetic features. We identified 344 unique drug targets in Pharmaprojects reported with a type 2 diabetes or diabetes mellitus indication, including 22 approved. We reviewed the list of approved drugs and eliminated those where there were questions around the relevance of the drug or target to T2D (*AKR1B1*, *AR*, *DRD1*, *HMGCR*, *IGF1R*, *LPL, SLC5A1*). Because Pharmaprojects ordinarily specifies the receptor as target for protein or peptide replacement therapies, we also remapped the minority of programs where the ligand, rather than receptor, had been listed as target (changing *INS* to *INSR*, *GCG* to *GCGR*) To assess the proportion of of programs with genetic support, we first grouped by drug and selected just one target, preferring the target with the earliest genetic support (OMIM, then established GWAS, then novel GWAS, then none). Next we grouped by target and selected its highest phase reached. Finally, we grouped by highest phase reached and counted the number of unique targets.

Universe of possible genetically supported gene-indication pairs. In all of our analyses, targets are defined as human gene symbols, but we use the term gene-indication pair (G-I) to refer to possible genes that one might attempt to target with a drug, and target-indication pair (T-I) to refer to genes that are the targets of actual drug candidates in development. To enumerate the space of possible G-I pairs, we multiplied the N=774 Pharmaprojects indications

considered here by the "universe" of N=19,338 protein-coding genes, yielding a space of N= 14,967,612 possible G-I pairs. Of these, N=103,688 (0.69%) qualify as having genetic support per our criteria. A total of 17,475 T-I pairs have reached at least Phase I in an active or historical program, of which 1,189 (6.8%) are genetically supported. This represents an enrichment compared to random chance (OR = 10.7, $P < 1.0e-15$, Fisher exact test), but in absolute terms, only 1.1% of genetically supported G-I pairs have been pursued. A genetically supported G-I pair may be less likely to attract drug development interest if the indication already has many other potential targets, and/or if the indication is but the second-most similar to the gene's associated trait. Removing associations with many GWAS hits and restricting to the single most similar indication left a space of 32,718 possible genetically supported G-I pairs, 727 (2.2%) of which had been pursued. This small percentage might yet be perceived to reflect competitive saturation, if the vast majority of indications are undevelopable and/or the vast majority of targets are undruggable. We therefore asked what proportion of genetically supported G-I pairs had been developed to at least Phase I, as a function of therapy area cross-tabulated against Open Targets predicted tractability status or membership in canonically "druggable" protein families, using families from ref. [15] as well as UniProt pkinfam for kinases[24]. We also grouped at the level of gene, rather than G-I pair (Figure S5).

Druggability and protein families. Antibody and small molecule druggability status was taken from Open Targets[25]. For antibody tractability, Clinical Precedence, Predicted Tractable – High Confidence, and Predicted Tractable – Medium to Low Confidence were included. For small molecules, Clinical Precedence, Discovery Precedence, and Predicted Tractable were included. Protein families were from sources described previously[15], plus the pkinfam kinase list from UniProt[24]. To make these lists non-overlapping, genes that were both kinases and also either enzymes, ion channels, or nuclear receptors were considered to be kinases only.

Statistics. Analyses utilized custom scripts in R 4.2.0. For binomial proportions P(G) and P(S), error bars are Wilson 95% confidence intervals. RS at each phase is defined as a risk ratio $(X_{yes}/N_{yes})/(X_{total}/N_{total})$ with Wilson 95% confidence intervals, while RS for phase I–launch is defined as a product of the three phase-wise risk ratios, with Katz 95% confidence intervals. Effect of continuous variables on probability of launch were assessed using logistic regression. Differences in RS between therapy areas were tested using the Cochran-Mantel-Haenszel chi-square test (cmh.test from the R lawstat package). Pipeline progression of drug-indication pairs conditioned on the highest phase reached by a drug was modeled using an ordinal logit model (polr with Hess=TRUE from the R MASS package). Correlations across therapy areas were tested by weighted Pearson's correlation (wtd.cor from the R weights package); to control for the amount of data available in each therapy area, the number of genetically supported T-I pairs having reached at least phase I used as the weight. Enrichments of T-I pairs in the utilization analysis were tested using Fisher's exact test. All statistical tests were two-sided.

Source code availability and data availability. An analytical dataset and source code will be made available at https://github.com/ericminikel/genetic_support/ and will be sufficient to reproduce all figures and statistics herein.

REFERENCES

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May;47:20–33. PMID: 26928437

2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014 Jan;32(1):40–51. PMID: 24406927

3. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019 01;20(2):273–286. PMCID: PMC6409418

4. Thomas D, Chancellor D, Micklus A, LaFever S, Hay M, Chaudhuri S, Bowden R, Lo AW. Clinical Development Success Rates and Contributing Factors 2011–2020 [Internet]. 2021 p. 34. Available from: https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf

5. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. Nat Genet. 2015 Aug;47(8):856–860. PMID: 26121088

6. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013 Aug;12(8):581–594. PMID: 23868113

7. Musunuru K, Kathiresan S. Genetics of Common, Complex Coronary Artery Disease. Cell. 2019 Mar 21;177(1):132–145. PMID: 30901535

8. Carss KJ, Deaton AM, Del Rio-Espinola A, Diogo D, Fielden M, Kulkarni DA, Moggs J, Newham P, Nelson MR, Sistare FD, Ward LD, Yuan J. Using human genetics to improve safety assessment of therapeutics. Nat Rev Drug Discov. 2023 Feb;22(2):145–162. PMID: 36261593

9. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012 Jan 13;90(1):7–24. PMCID: PMC3257326

10. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15(12):e1008489. PMCID: PMC6907751

11. Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. Nat Rev Genet. 2021 Oct;22(10):658–671. PMID: 34302145

12. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, Huffman JE, Assimes TL, Lorenz K, Zhu X, Hilliard AT, Judy RL, Huang J, Lee KM, Klarin D, Pyarajan S, Danesh J, Melander O, Rasheed A, Mallick NH, Hameed S, Qureshi IH, Afzal MN, Malik U, Jalal A, Abbas S, Sheng X, Gao L, Kaestner KH, Susztak K, Sun YV, DuVall SL, Cho K, Lee JS, Gaziano JM, Phillips LS, Meigs JB, Reaven PD, Wilson PW, Edwards TL, Rader DJ, Damrauer SM, O'Donnell CJ, Tsao PS, HPAP Consortium, Regeneron Genetics Center, VA Million Veteran Program, Chang KM, Voight BF, Saleheen D. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat Genet. 2020 Jun 15; PMID: 32541925

13. Lommatzsch M, Brusselle GG, Canonica GW, Jackson DJ, Nair P, Buhl R, Virchow JC. Disease-modifying anti-asthmatic drugs. Lancet. 2022 Apr 23;399(10335):1664–1668. PMID: 35461560

14. Mortberg MA, Vallabh SM, Minikel EV. Disease stages and therapeutic hypotheses in two decades of neurodegenerative disease clinical trials. Sci Rep. 2022 Oct 21;12(1):17708. PMCID: PMC9587287

15. Minikel EV, Karczewski KJ, Martin HC, Cummings BB, Whiffin N, Rhodes D, Alföldi J, Trembath RC, van Heel DA, Daly MJ, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Schreiber SL, MacArthur DG. Evaluating drug targets through human loss-of-function genetic variation. Nature. 2020 May;581(7809):459–464. PMCID: PMC7272226

16. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177–1186. PMCID: PMC5536862

17. Scannell JW, Bosley J, Hickman JA, Dawson GR, Truebel H, Ferreira GS, Richards D, Treherne JM. Predictive validity in drug discovery: what it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022 Dec;21(12):915–931. PMID: 36195754

18. Citeline I Data Analysis | Pharma Intelligence [Internet]. [cited 2023 Apr 13]. Available from: https://pharmaintelligence.informa.com/products-and-services/clinical-planning/citeline

19. Painter JL. Toward automating an inference model on unstructured terminologies: OXMIS case study. Adv Exp Med Biol. 2010;680:645–651. PMID: 20865550

20. Guo C, Sieber KB, Esparza-Gordillo J, Hurle MR, Song K, Yeo AJ, Yerges-Armstrong LM, Johnson T, Nelson MR. Identification of putative effector genes across the GWAS Catalog using molecular quantitative trait loci from 68 tissues and cell types. bioRxiv. 2019 Jan 1;808444.

21. Karczewski KJ, Solomonson M, Chao KR, Goodrich JK, Tiao G, Lu W, Riley-Gillis BM, Tsai EA, Kim HI, Zheng X, Rahimov F, Esmaeeli S, Grundstad AJ, Reppell M, Waring J, Jacob H, Sexton D, Bronson PG, Chen X, Hu X, Goldstein JI, King D, Vittal C, Poterba T, Palmer DS, Churchhouse C, Howrigan DP, Zhou W, Watts NA, Nguyen K, Nguyen H, Mason C, Farnham C, Tolonen C, Gauthier LD, Gupta N, MacArthur DG, Rehm HL, Seed C, Philippakis AA, Daly MJ, Davis JW, Runz H, Miller MR, Neale BM. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genomics. 2022 Sep 14;2(9):100168.

22. Lin D. An information-theoretic definition of similarity. Icml. 1998. p. 296–304.

23. Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of artificial intelligence research. 1999;11:95–130.

24. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 04;45(D1):D158–D169. PMCID: PMC5210571

25. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, Lopez I, Miranda A, Cruz-Castillo C, Fumis L, Bernal-Llinares M, Tsukanov K, Cornu H, Tsirigos K, Razuvayevskaya O, Buniello A, Schwartzentruber J, Karim M, Ariano B, Martinez Osorio RE, Ferrer J, Ge X, Machlitt-Northen S, Gonzalez-Uriarte A, Saha S, Tirunagari S, Mehta C, Roldán-Romero JM, Horswell S, Young S, Ghoussaini M, Hulcoop DG, Dunham I, McDonagh EM. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023 Jan 6;51(D1):D1353–D1359. PMCID: PMC9825572

SUPPLEMENTARY FIGURES

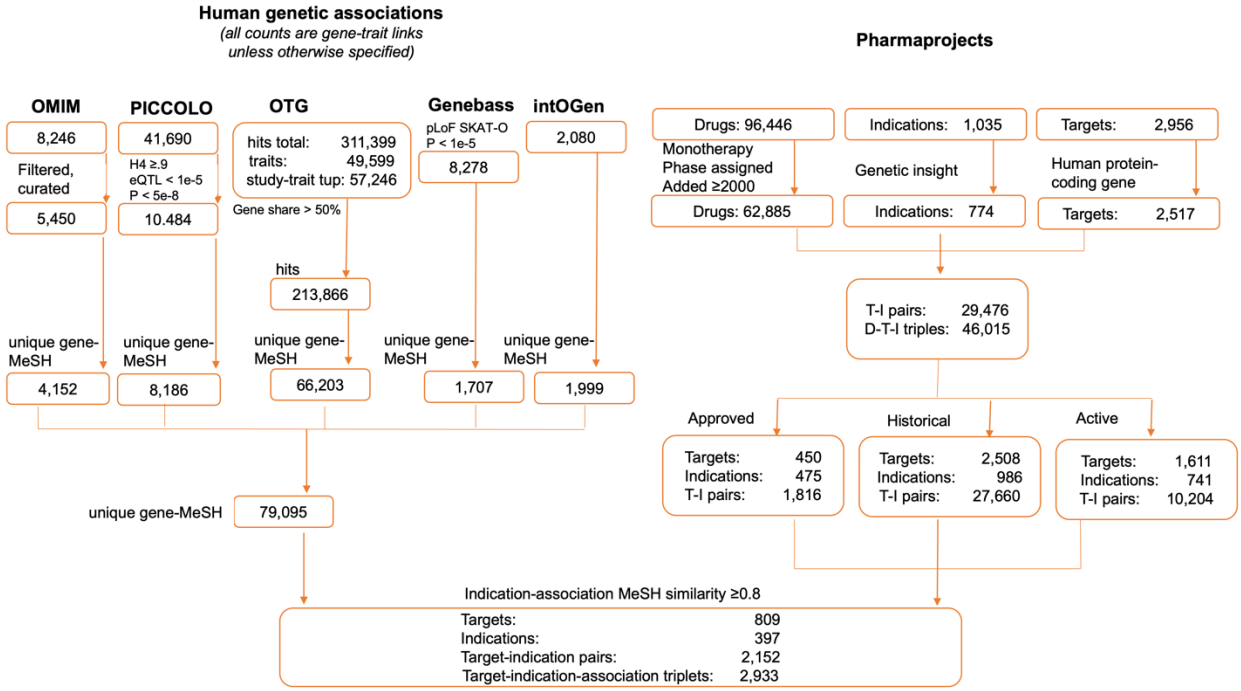


Figure S1. Data processing schematic. *Dataset size, filters, and join process for Pharmaprojects and human genetic evidence. Note that a drug can be assigned multiple targets, and can be approved for multiple indications. The entire analysis described herein has also been run restricted to only those drugs with exactly one target annotated (Figures S5-S7).*

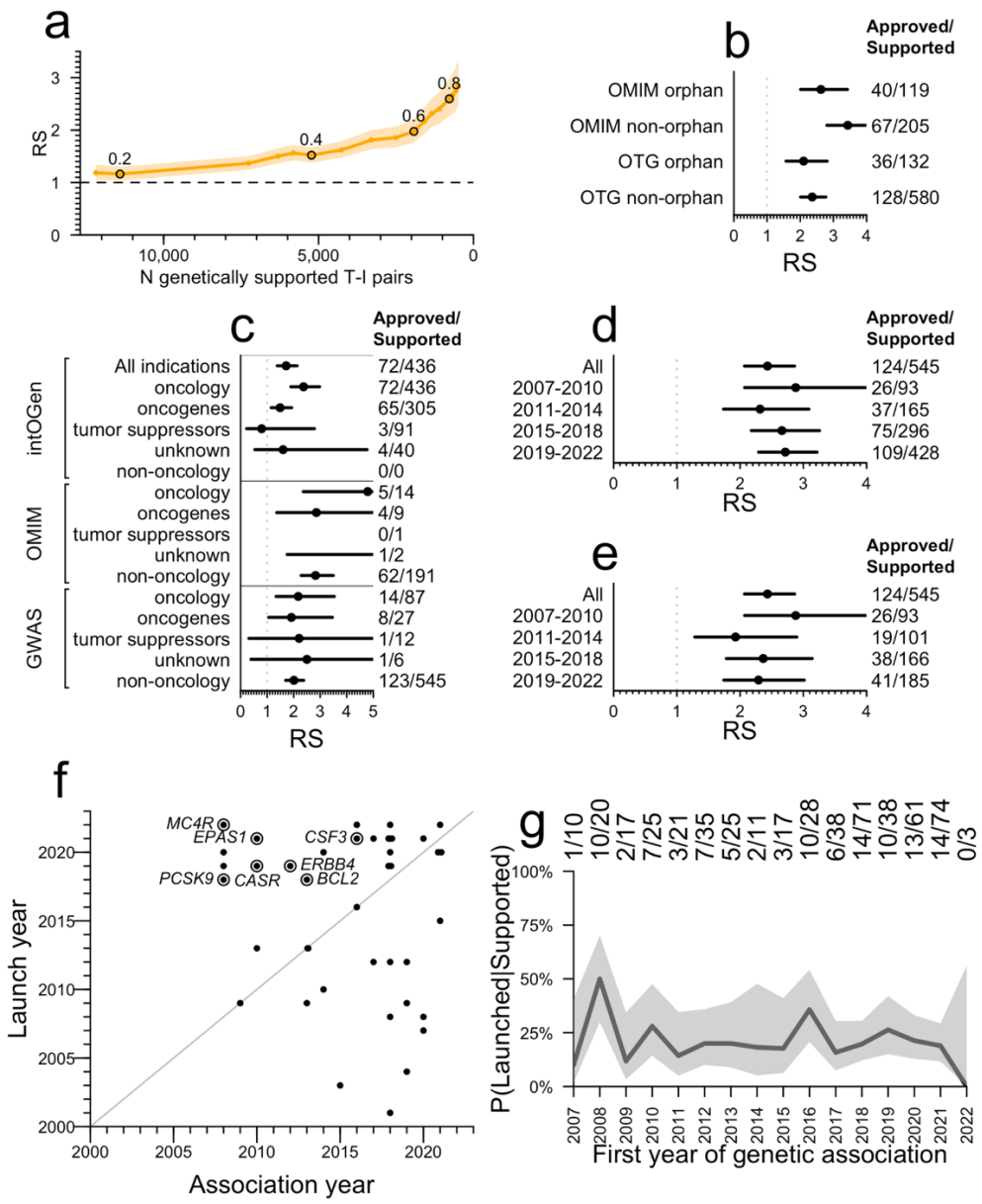


Figure S2. Further analysis of influence of characteristics of genetic associations on relative success. A) *Sensitivity of RS to the similarity threshold between the MeSH ID for the genetically associated trait and the MeSH ID for the clinically developed indication. The threshold is varied by units of 0.05 (labels) and the results are plotted as RS (y axis) versus number of genetically supported T-I pairs (x axis).* ***B)*** *Breakdown of OTG and OMIM RS values by whether any drug for each target-indication pair has had orphan status assigned.* ***C)*** *RS for somatic genetic evidence from IntOGen versus germline genetic evidence, for oncology and non-oncology indications.* ***D)*** *As for top panel of Figure 1D, but without removing replications or OMIM-supported T-I pairs.* ***E)*** *As for top panel of Figure 1D, removing replications but not removing OMIM-supported T-I pairs.* ***F)*** *Launched, genetically supported T-I pairs by year of launch (y axis) and year of first genetic association (x axis). Gene symbols are labeled for first approvals of targets with at least 5 years between association and launch.* ***G)*** *Proportion of T-I pairs supported by a GWAS Catalog association that are launched (versus phase I-III) as a function of the year of first genetic association.*

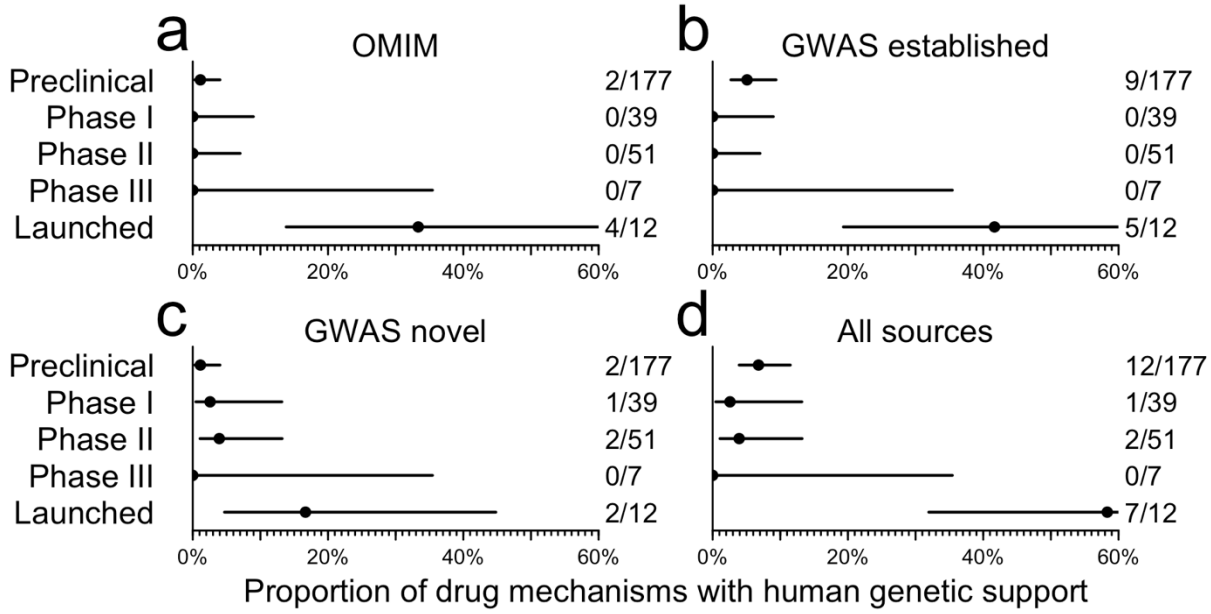


Figure S3. Proportion of type 2 diabetes drug targets with human genetic support by highest phase reached. *A) OMIM, b) established GWAS genes, c) novel GWAS genes, or d) any of the above.*

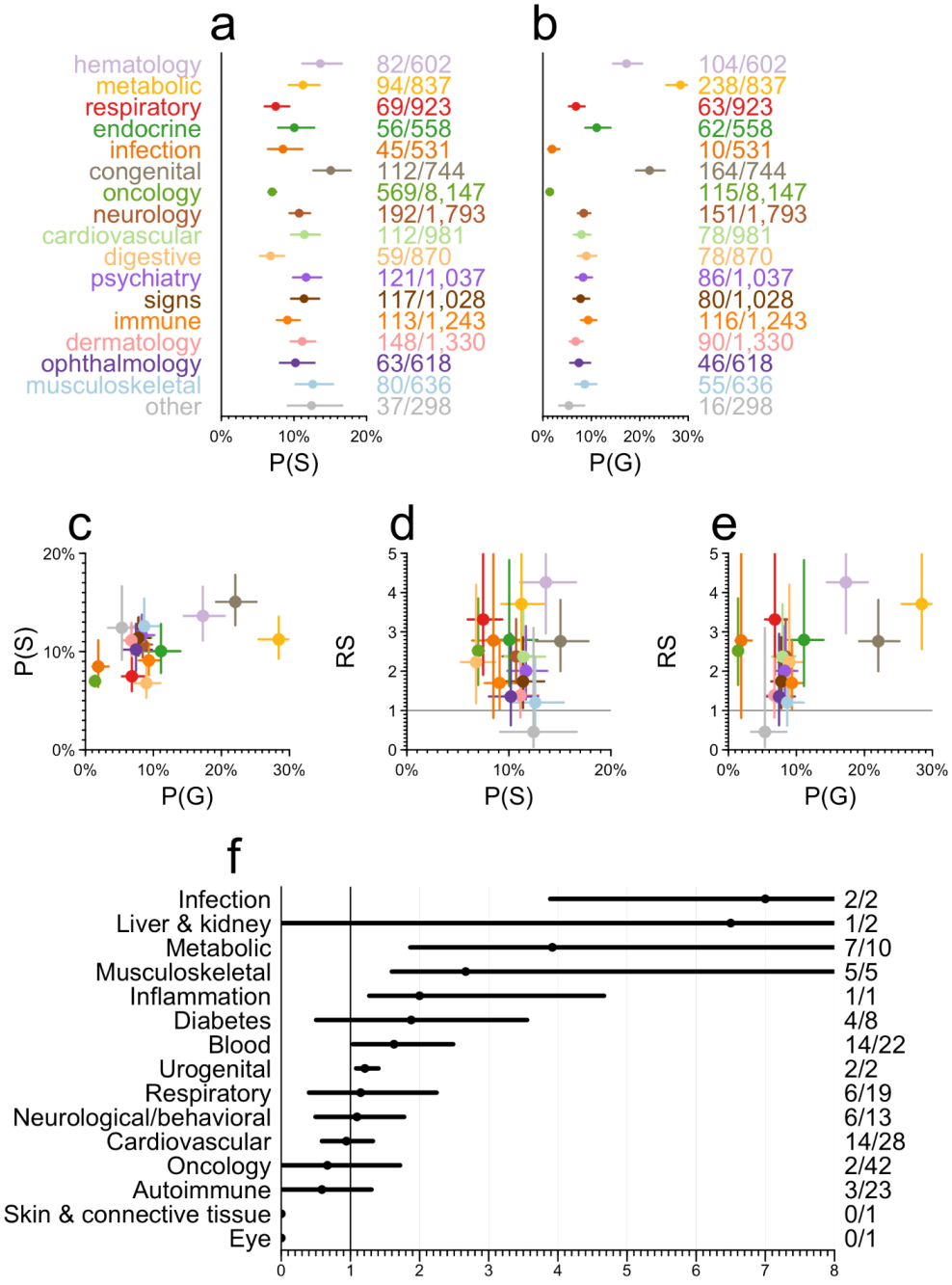
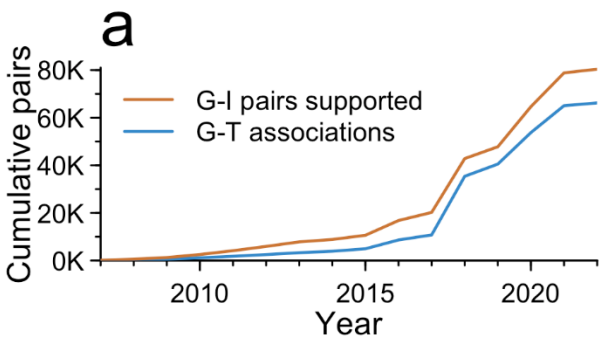


Figure S4. Further analyses of differences in relative success among therapy areas. A) *Probability of success, P(S), by therapy area, with Wilson 95% confidence intervals. Fractions at right show the number of launched T-I pairs (numerator) and number of T-I pairs reaching at least phase I (denominator).* ***B)*** *Probability of genetic support, P(G), by therapy area, with Wilson 95% confidence intervals. Fractions at right show the number of genetically supported T-I pairs reaching at least phase I (numerator) and total number of T-I pairs reaching at least phase I (denominator).* ***C)*** *P(S) vs. P(G),* ***D)*** *RS s. P(S), and* ***E)*** *RS vs. P(G) across therapy areas, with crosshairs representing 95% confidence intervals on both dimensions.* ***F)*** *Re-analysis of RS (x axis) broken down by therapy area using data from supplementary table 6 of Nelson et al. 2015.*

a



b

	all genes	predicted Ab tractable	predicted SM tractable	rhodopsin-like GPCRs	nuclear receptors	enzymes	ion channels	kinases
all areas	371/9913 **3.7%**	294/5061 **5.8%**	298/3066 **9.7%**	25/162 **15%**	11/35 **31%**	65/580 **11%**	22/205 **11%**	34/231 **15%**
hematology	36/2857 **1.3%**	29/1394 **2.1%**	32/968 **3.3%**	1/39 **2.6%**	0/13 **0%**	9/171 **5.3%**	0/33 **0%**	3/82 **3.7%**
metabolic	86/3668 **2.3%**	71/1919 **3.7%**	64/1237 **5.2%**	5/49 **10%**	5/20 **25%**	10/242 **4.1%**	1/52 **1.9%**	2/100 **2%**
respiratory	17/678 **2.5%**	16/360 **4.4%**	11/234 **4.7%**	2/6 **33%**	0/4 **0%**	2/41 **4.9%**	1/18 **5.6%**	0/10 **0%**
endocrine	29/1957 **1.5%**	26/977 **2.7%**	25/639 **3.9%**	5/33 **15%**	2/22 **9.1%**	1/132 **0.76%**	1/40 **2.5%**	3/52 **5.8%**
infection	4/211 **1.9%**	4/126 **3.2%**	2/75 **2.7%**	0/4 **0%**	0/2 **0%**	0/11 **0%**	0/6 **0%**	0/2 **0%**
congenital	84/3319 **2.5%**	63/1797 **3.5%**	58/1213 **4.8%**	2/37 **5.4%**	2/18 **11%**	15/221 **6.8%**	1/73 **1.4%**	2/82 **2.4%**
oncology	108/2025 **5.3%**	77/928 **8.3%**	95/705 **13%**	3/26 **12%**	3/13 **23%**	24/128 **19%**	0/33 **0%**	27/75 **36%**
neurology	77/2669 **2.9%**	64/1461 **4.4%**	59/904 **6.5%**	7/44 **16%**	0/8 **0%**	10/159 **6.3%**	9/86 **10%**	3/63 **4.8%**
cardiovascular	34/1762 **1.9%**	32/932 **3.4%**	31/590 **5.3%**	2/19 **11%**	1/11 **9.1%**	10/129 **7.8%**	7/48 **15%**	1/59 **1.7%**
digestive	29/1546 **1.9%**	23/837 **2.7%**	20/525 **3.8%**	2/31 **6.5%**	2/12 **17%**	6/90 **6.7%**	1/21 **4.8%**	2/50 **4%**
psychiatry	32/1565 **2%**	30/857 **3.5%**	31/491 **6.3%**	9/34 **26%**	0/9 **0%**	4/96 **4.2%**	6/54 **11%**	0/48 **0%**
signs	45/2558 **1.8%**	38/1359 **2.8%**	38/812 **4.7%**	7/51 **14%**	4/20 **20%**	4/150 **2.7%**	8/69 **12%**	3/69 **4.3%**
immune	40/1430 **2.8%**	37/772 **4.8%**	25/492 **5.1%**	3/22 **14%**	0/4 **0%**	7/82 **8.5%**	0/19 **0%**	2/47 **4.3%**
dermatology	38/1644 **2.3%**	34/898 **3.8%**	22/528 **4.2%**	1/17 **5.9%**	0/7 **0%**	6/109 **5.5%**	0/25 **0%**	3/44 **6.8%**
ophthalmology	22/1517 **1.5%**	16/855 **1.9%**	14/450 **3.1%**	1/15 **6.7%**	1/12 **8.3%**	3/84 **3.6%**	2/42 **4.8%**	0/39 **0%**
musculoskeletal	29/1408 **2.1%**	26/765 **3.4%**	21/462 **4.5%**	1/16 **6.2%**	2/6 **33%**	2/85 **2.4%**	1/25 **4%**	4/45 **8.9%**
other	7/1126 **0.62%**	5/570 **0.88%**	6/351 **1.7%**	0/19 **0%**	1/7 **14%**	0/67 **0%**	0/26 **0%**	1/28 **3.6%**

Figure S5. Level of utilization of genetic support among targets. *As for Figure 3, but grouped by target instead of T-I pair. Thus, the denominator for each cell is the number of targets with at least one genetically supported indication, and each target counts in the numerator if at least one genetically supported indication has reached phase I.*

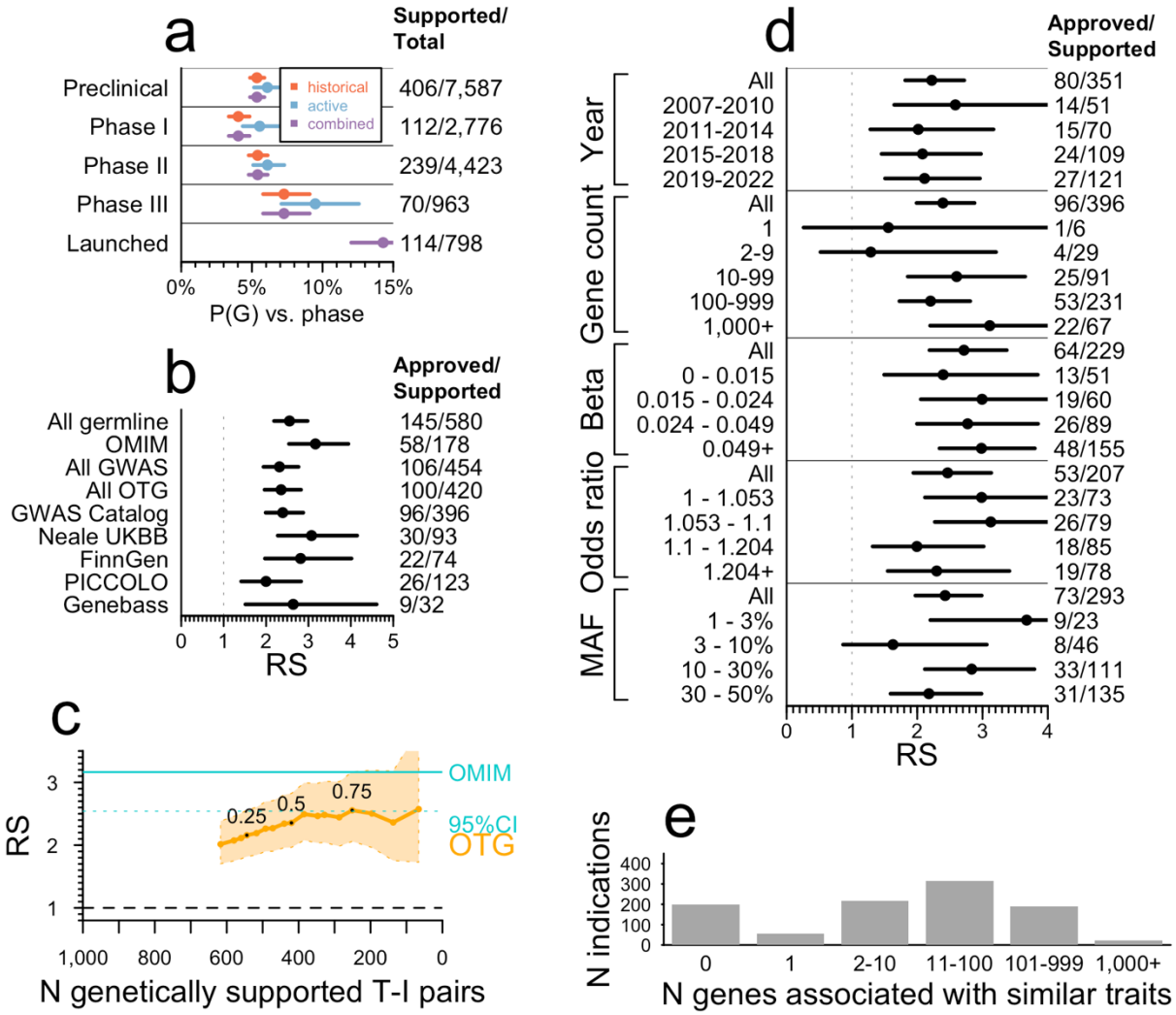


Figure S6. Figure 1 restricted to drugs with one target only.

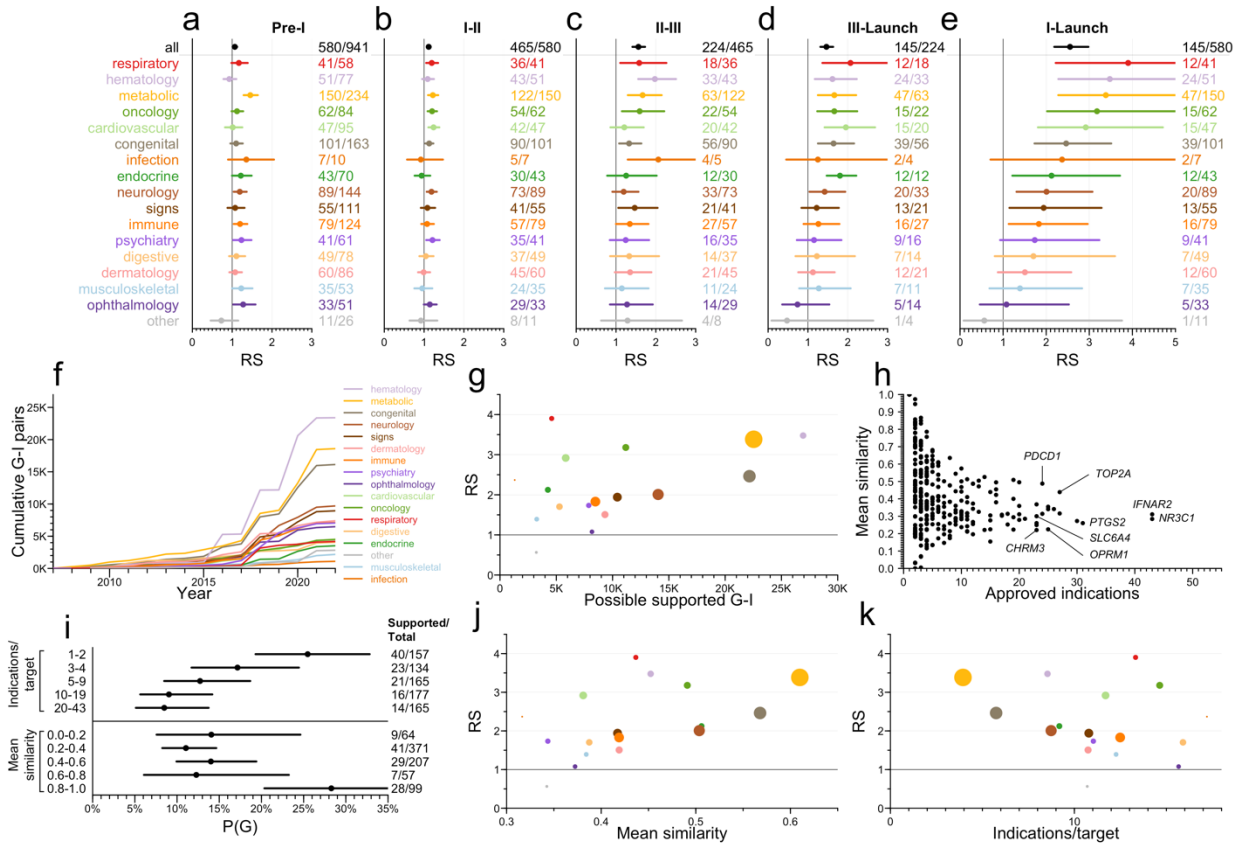


Figure S7. Figure 2 restricted to drugs with one target only.

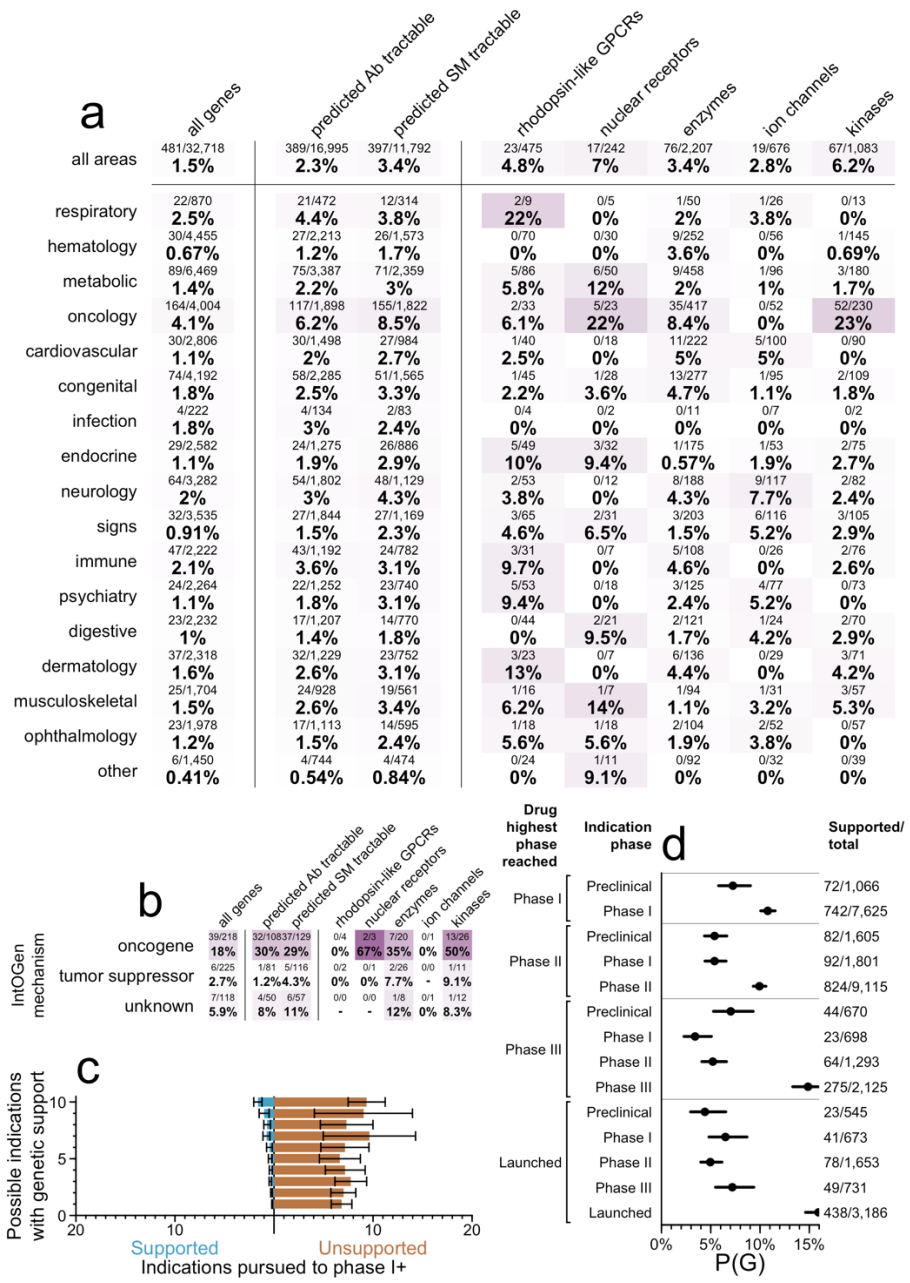


Figure S8. Figure 3 restricted to drugs with one target only.